



## Local binary regression with spherical predictors

Marco Di Marzio <sup>a,\*</sup>, Stefania Fensore <sup>a</sup>, Agnese Panzera <sup>b</sup>, Charles C. Taylor <sup>c</sup>



<sup>a</sup> DSFPEQ, Università di Chieti-Pescara, viale Pindaro 42, 65127 Pescara, Italy

<sup>b</sup> DiSIA, Università di Firenze, viale Morgagni 59, 50134 Florence, Italy

<sup>c</sup> Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

### ARTICLE INFO

#### Article history:

Available online 26 July 2018

#### MSC:

62H11

62H20

#### Keywords:

Local likelihood

Spherical kernels

Tangent-normal decomposition

### ABSTRACT

We discuss local regression estimators when the predictor lies on the  $d$ -dimensional sphere and the response is binary. Despite Di Marzio et al. (2019), who introduce spherical kernel density classification, we build on the theory of local polynomial regression and local likelihood. Simulations and a real-data application illustrate the effectiveness of the proposals.

© 2018 Elsevier B.V. All rights reserved.

## 1. Introduction

Data lying on the unit hypersphere embedded in  $\mathbb{R}^d$ ,  $d \geq 2$ , arise in many scientific fields. They are typically referred as *directional* or *spherical* data. Classical examples, when  $d = 2$ , are directions of winds and marine currents, and directions of flight of birds from a point of release. Also, locations on the surface of the *ordinary* sphere ( $d = 3$ ) are ubiquitous in Earth and planetary sciences.

The non-linear nature of the hypersphere sets apart directional statistics from standard methods, which are typically designed for *linear* data. However, in the last few decades *directional statistics* has greatly evolved, and now directional counterparts of many classical statistical methods exist. A classical account of directional statistics is Mardia and Jupp (2008), and very recent ones are Ley and Verdebout (2017, 2018).

Kernel-based methods for regression estimation when the response is a linear variable and the predictor has a directional nature have been recently studied. Indeed, the absence of a *boundary* on a spherical domain makes smoothing methods – which typically suffer from boundary bias – well-suited for analysing directional data. In particular, the local polynomial regression for linear response has been studied by Di Marzio et al. (2009) in the case of circular predictors, and by Di Marzio et al. (2014) in the case of a general  $d$ -dimensional spherical predictor, as an intermediate step in the spherical–spherical regression estimation. Then, this topic has been also studied by García-Portugués et al. (2016) in the context of goodness-of-fit tests.

Conversely, the special case of a binary response and a directional predictor by means of nonparametric regression methods seems to be unexplored, while for a parametric approach see Fernandes and Cardoso (2016) and references therein. The binary regression problem, apart from being of interest *per se*, is also useful for classification purposes. Nonparametric methods for classification of directional data, based on kernel estimation of spherical densities, have been proposed by Di Marzio et al. (2019).

\* Corresponding author.

E-mail addresses: [marco.dimarzio@unich.it](mailto:marco.dimarzio@unich.it) (M. Di Marzio), [stefania.fensore@unich.it](mailto:stefania.fensore@unich.it) (S. Fensore), [agnese.panzera@unifi.it](mailto:agnese.panzera@unifi.it) (A. Panzera), [charles@maths.leeds.ac.uk](mailto:charles@maths.leeds.ac.uk) (C.C. Taylor).

In the Euclidean setting, kernel-based estimators of the binary regression with a *linear* predictor have been studied by [Fan et al. \(1995\)](#) and [Signorini and Jones \(2004\)](#), who provided asymptotic properties of various versions of the estimators. The discussed methods essentially rely on local polynomial regression and a local likelihood approach. We adapt these theories to the spherical setting. Finally, as a close reference, we recall that a local-likelihood-based approach has been also investigated by [Di Marzio et al. \(2018\)](#) in the different context of the estimation of densities defined on the  $d$ -dimensional torus.

The paper is organized as follows. In Section 2 we recall a Taylor-like polynomial to approximate functions having the unit hypersphere as their domain. In Section 3 we discuss the adaptation of the theory of local polynomial regression with a directional predictor to the binary response case, while, in Section 4 we propose the nonparametric estimation using a locally weighted likelihood objective function. Finally, Section 5 collects some simulation examples and a real-data application.

## 2. Series expansion for functions on the sphere

Let  $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$  denote the unit hypersphere embedded in  $\mathbb{R}^d$ ,  $d \geq 2$ . The *tangent-normal* decomposition provides a parametrization of a point on  $\mathbb{S}^{d-1}$  as follows. For fixed  $\mathbf{x} \in \mathbb{S}^{d-1}$ , any vector  $\mathbf{u} \in \mathbb{S}^{d-1}$  can be expressed as  $\mathbf{u}(\xi, \theta) = \mathbf{x} \cos(\theta) + \xi \sin(\theta)$ , where  $\theta$  is the angle between  $\mathbf{u}$  and  $\mathbf{x}$ , and  $\xi$  is a unit vector orthogonal to  $\mathbf{x}$ . Now, letting  $\mu_d$  denote the Lebesgue measure of  $\mathbb{S}^d$ , with  $\omega_d = \mu_d(\mathbb{S}^d) = \frac{2\pi^{(d+1)/2}}{\Gamma((d+1)/2)}$ , and setting  $\mathbb{T}_{\mathbf{x}} = \{\xi \in \mathbb{S}^{d-1} : \xi \perp \mathbf{x}\}$ , for a real-valued function  $g$  defined on  $\mathbb{S}^{d-1}$ , the integration formula corresponding to the above parametrization is

$$\int_{\mathbb{S}^{d-1}} g(\mathbf{u}) d\mu_{d-1}(\mathbf{u}) = \int_0^\pi \sin^{d-2}(\theta) d\theta \int_{\mathbb{T}_{\mathbf{x}}} g(\mathbf{u}(\xi, \theta)) d\mu_{d-2}(\xi).$$

Moreover, letting  $\bar{g}(\mathbf{x}) := g(\mathbf{x}/\|\mathbf{x}\|)$  be the homogeneous extension of  $g$  to  $\mathbb{R}^d \setminus \{\mathbf{0}_d\}$ , with  $\mathbf{0}_d$  being the  $d$ -dimensional zero vector, we have that

$$\left. \frac{\partial^\ell}{\partial \theta^\ell} g(\mathbf{u}(\xi, \theta)) \right|_{\theta=0} = \mathcal{D}_\xi^{(\ell)} \bar{g}(\mathbf{x}),$$

where  $\mathcal{D}_\xi^{(\ell)} \bar{g}(\mathbf{x})$  is the *directional derivative* of order  $\ell$  of  $\bar{g}$  at  $\mathbf{x}$  in the direction of  $\xi$ . Clearly  $\mathcal{D}_\xi^{(0)} \bar{g}(\mathbf{x}) = g(\mathbf{x})$ , while, letting  $\nabla_{\bar{g}}^\ell(\mathbf{x})$  be the matrix of the derivatives of total order  $\ell$  of  $\bar{g}$  at  $\mathbf{x}$ , one has

$$\mathcal{D}_\xi^{(\ell)} \bar{g}(\mathbf{x}) = \xi' \nabla_{\bar{g}}^\ell(\mathbf{x}) \xi^{\otimes(\ell-1)},$$

where  $\mathbf{a}^{\otimes \ell}$  stands for the Kroneckerian power of order  $\ell$  of a vector  $\mathbf{a}$ . For example, we have  $\mathcal{D}_\xi^{(1)} \bar{g}(\mathbf{x}) = \xi' \nabla_{\bar{g}}(\mathbf{x})$  and  $\mathcal{D}_\xi^{(2)} \bar{g}(\mathbf{x}) = \xi' \nabla_{\bar{g}}^2(\mathbf{x}) \xi$ , with  $\nabla_{\bar{g}}^1(\mathbf{x})$  and  $\nabla_{\bar{g}}^2(\mathbf{x})$  respectively being the gradient vector and the Hessian matrix of  $\bar{g}$  at  $\mathbf{x}$ , while  $\mathcal{D}_\xi^{(3)} \bar{g}(\mathbf{x}) = \xi' \nabla_{\bar{g}}^3(\mathbf{x}) \xi \otimes \xi$ , with  $\mathbf{a} \otimes \mathbf{a}$  being the Kroneckerian product of the vector  $\mathbf{a}$  by itself.

Now, under suitable continuity assumptions, a Taylor-like expansion of a real valued function  $g$  defined on  $\mathbb{S}^{d-1}$  can be provided. Specifically, by assuming the continuity of  $\nabla_{\bar{g}}^\ell(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{S}^{d-1}$ , for  $\ell \in (1, \dots, p)$ , a  $p$ th-order series expansion of  $g$  around  $\mathbf{x}$  yields

$$g(\mathbf{u}) \approx g(\mathbf{x}) + \sum_{\ell=1}^p \frac{\theta^\ell}{\ell!} \mathcal{D}_\xi^{(\ell)} \bar{g}(\mathbf{x}) = g(\mathbf{x}) + \sum_{\ell=1}^p \frac{\theta^\ell}{\ell!} \xi' \nabla_{\bar{g}(\mathbf{x})}^\ell \xi^{\otimes(\ell-1)}. \quad (1)$$

The above expansion has been employed for deriving the asymptotic properties of kernel estimators for spherical densities by [Hall et al. \(1987\)](#) and [Klemela \(2000\)](#), to obtain a component-wise local approximation of spherical-spherical regression by [Di Marzio et al. \(2014\)](#), and to approximate the entries of skew-symmetric matrices and define rotations for spherical regression by [Di Marzio et al. \(2018\)](#).

## 3. Local polynomial binary regression

Let  $(\mathbf{X}, Y)$  be a  $\mathbb{S}^{d-1} \times \{0, 1\}$ -valued random variable, and set  $\lambda(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$ . If independent copies  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  of  $(\mathbf{X}, Y)$  are observed, by ignoring the binary nature of  $\mathbf{Y}$ , a naive nonparametric estimation of  $\lambda(\mathbf{x})$  can be performed by using the local polynomial estimators with real-valued response and spherical predictor, which have been studied by [Di Marzio et al. \(2014\)](#). In particular, following this approach, the regression function at  $\mathbf{x}_i$  is approximated by a suitable  $p$ th degree polynomial around  $\mathbf{x} \in \mathbb{S}^{d-1}$ , and a local estimator of  $\lambda(\mathbf{x})$  is defined as the solution (for the zero order coefficient) of the minimization of a weighted  $L_2$  distance between the  $Y_i$ s and the approximating polynomial. Different values of  $p$  give different estimators. Formally, by using expansion (1), a  $p$ th degree local polynomial estimator of  $\lambda$  at  $\mathbf{x} \in \mathbb{S}^{d-1}$ , say  $\hat{\lambda}(\mathbf{x}; p)$ , can be defined as the solution for  $\beta_0$  of

$$\operatorname{argmin}_{\{\beta_0, \beta_1, \dots, \beta_p\}} \sum_{i=1}^n \left\{ Y_i - \beta_0 - \sum_{\ell=1}^p \frac{\theta_i^\ell}{\ell} \xi_i' \beta_\ell \xi_i^{\otimes(\ell-1)} \right\}^2 K_k(\mathbf{x}' \mathbf{X}_i), \quad (2)$$

where  $\theta_i$  is the angle between  $\mathbf{X}_i$  and  $\mathbf{x}$ , and the weight  $K_\kappa$  is a *spherical kernel*. A spherical kernel can be essentially defined as a unimodal density having  $\mathbb{S}^{d-1}$  as its support, with rotational symmetry about its mean direction  $\boldsymbol{\mu} = (0, \dots, 0, 1)$ , and concentration parameter  $\kappa > 0$  such that as  $\kappa$  increases  $K_\kappa$  concentrates around  $\boldsymbol{\mu}$ . In Eq. (2)  $K_\kappa(\mathbf{x}'\mathbf{X}_i)$  is a spherical kernel with mean direction  $\mathbf{X}_i$ , and evaluated at  $\mathbf{x}$ . This weight function emphasizes the contribution of the observations  $\mathbf{X}_i$ s which are closer to the estimation point  $\mathbf{x}$ . Kernels of this form have been used by Hall et al. (1987) for density estimation on the sphere and by Di Marzio et al. (2014) and Di Marzio et al. (2018) for spherical-spherical regression estimation.

Now, when  $p = 0$ , the solution for  $\beta_0$  leads to the *local constant* estimator

$$\hat{\lambda}(\mathbf{x}; 0) = \frac{\sum_{i=1}^n Y_i K_\kappa(\mathbf{x}'\mathbf{X}_i)}{\sum_{i=1}^n K_\kappa(\mathbf{x}'\mathbf{X}_i)}, \quad (3)$$

while, when  $p = 1$ , the unique solution for  $\beta_0$  of the above least squares problem under a suitable constraint (see Di Marzio et al., 2014 for details) is

$$\hat{\lambda}(\mathbf{x}; 1) = \sum_{i=1}^n W_\kappa(\mathbf{x}'\mathbf{X}_i) Y_i,$$

where  $W_\kappa(\mathbf{x}'\mathbf{X}_i) = \mathbf{x}' \left\{ \sum_{j=1}^n K_\kappa(\mathbf{x}'\mathbf{X}_j)(\mathbf{x} + \theta_j \boldsymbol{\xi}_j)(\mathbf{x} + \theta_j \boldsymbol{\xi}_j)' \right\}^{-1} (\mathbf{x} + \theta_i \boldsymbol{\xi}_i) K_\kappa(\mathbf{x}'\mathbf{X}_i)$ .

Now, in order to discuss the asymptotic properties of the estimators, we need to recall the *spherical counterparts* of the  $j$ th moment,  $j \in \mathbb{N}$ , and the roughness of a Euclidean kernel, which, for a kernel  $K_\kappa$  respectively are

$$b_j(\kappa) = \omega_{d-2} \int_0^\pi K_\kappa(\cos(\theta)) \theta^j \sin^{d-2}(\theta) d\theta, \quad v_0(\kappa) = \omega_{d-2} \int_0^\pi K_\kappa^2(\cos(\theta)) \sin^{d-2}(\theta) d\theta.$$

Let  $\text{Tr}(\mathbf{A})$  denote the trace of the matrix  $\mathbf{A}$ , and use  $f$  to denote the common density of the  $\mathbf{X}_i$ s. Then, for the cases  $p = 0$  and  $p = 1$ , by using respectively results in Theorem 1 and Theorem 2 in Di Marzio et al. (2014), we obtain the following result.

**Result 1.** Given the  $\mathbb{S}^{d-1} \times \{0, 1\}$ -valued random sample  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ , consider estimator  $\hat{\lambda}(\mathbf{x}; p)$ ,  $\mathbf{x} \in \mathbb{S}^{d-1}$ . If

- (i)  $K_\kappa$  is a spherical kernel such that as  $n$  increases  $b_2(\kappa)$  and  $v_0(\kappa)/n$  both go to 0, and for  $j > 2$ ,  $b_j(K_\kappa) = o(b_2(\kappa))$ ;
- (ii)  $f(\mathbf{x}) > 0$  and all the entries of  $\nabla_{\bar{f}}(\mathbf{x})$ ,  $\nabla_{\bar{\lambda}}(\mathbf{x})$ , and  $\nabla_{\bar{\lambda}}^2(\mathbf{x})$  are continuous,

then

$$\begin{aligned} \mathbb{E}[\hat{\lambda}(\mathbf{x}; 0)] - \lambda(\mathbf{x}) &= \frac{b_2(\kappa)}{2(d-1)} \left( \text{Tr} \left\{ \nabla_{\bar{\lambda}}^2(\mathbf{x}) \right\} + \frac{2\nabla'_{\bar{\lambda}}(\mathbf{x})\nabla_{\bar{f}}(\mathbf{x})}{f(\mathbf{x})} \right) + o(b_2(\kappa)), \\ \mathbb{E}[\hat{\lambda}(\mathbf{x}; 1)] - \lambda(\mathbf{x}) &= \frac{b_2(\kappa)}{2(d-1)} \text{Tr} \left\{ \nabla_{\bar{\lambda}}^2(\mathbf{x}) \right\} + o(b_2(\kappa)), \end{aligned}$$

and, for both  $p = 0$  and  $p = 1$ ,

$$\text{Var}[\hat{\lambda}(\mathbf{x}; p)] = \frac{v_0(\kappa)}{n} \frac{\lambda(\mathbf{x})(1 - \lambda(\mathbf{x}))}{f(\mathbf{x})} + o\left(\frac{v_0(\kappa)}{n}\right).$$

**Remark 1.** Recently, García-Portugués et al. (2016) proposed a different series expansion of the regression function with linear response and directional predictor, which generalizes the proposal of Di Marzio et al. (2009) in the circular case when  $p = 1$ . The optimization of the corresponding  $L_2$  loss leads to a projected local linear estimator which shares the asymptotic properties of the local linear estimator of Di Marzio et al. (2014).

An optimal smoothing degree would minimize the asymptotic mean-squared error of  $\hat{\lambda}(\mathbf{x}; p)$ , which is the sum of the leading terms of the asymptotic squared bias and the asymptotic variance. Notice that the dependence of asymptotic bias and variance on the concentration parameter cannot be generalized with respect to the kernel, because it is not a scale factor. However it is important to detail the case of a von Mises–Fisher kernel, which can be regarded as the spherical counterpart of the Gaussian one. It is defined as  $K_\kappa(\mathbf{x}'\boldsymbol{\mu}) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} \mathcal{J}_{d/2-1}(\kappa)} \exp(\kappa \mathbf{x}'\boldsymbol{\mu})$ , with  $\mathcal{J}_u(\cdot)$  being the modified Bessel function of the first kind and order  $u$ . When  $\kappa$  is big enough, and  $j \in \mathbb{N}$ , the following approximations of  $b_j(\kappa)$  and  $v_0(\kappa)$  hold:

$$b_j(\kappa) \sim \frac{2^{j/2} \Gamma((d+j-1)/2)}{\kappa^{j/2} \Gamma((d-1)/2)}, \quad \text{and} \quad v_0(\kappa) \sim \frac{\kappa^{(d-1)/2}}{2^{d-1} \pi^{(d-1)/2}}.$$

As a consequence, when  $K_\kappa$  is a von Mises–Fisher kernel, the asymptotic bias and the asymptotic variance, for both  $p = 0$  and  $p = 1$ , are

$$\mathbb{E}[\hat{\lambda}(\mathbf{x}; p)] - \lambda(\mathbf{x}) = O\left(\frac{1}{\kappa}\right), \quad \text{and} \quad \text{Var}[\hat{\lambda}(\mathbf{x}; p)] = O\left(\frac{\kappa^{(d-1)/2}}{n}\right).$$

Then for both local constant and local linear estimators, the value of  $\kappa$  which minimizes the asymptotic mean squared error is  $O(n^{2/(d+3)})$  and gives a convergence rate of magnitude  $O(n^{-4/(d+3)})$ . This is the same rate attained by single bandwidth local constant and local linear estimators of a real-valued regression function defined on  $\mathbb{R}^{d-1}$ , when a second-order kernel is employed.

#### 4. Local logistic regression via likelihood

The approach discussed in the previous section does not produce *bona-fide* estimates when the polynomial degree is greater than 0. Despite the fact that a truncation could be used for exploratory data analysis, the subsequent lack of differentiability may be a serious issue. To take into account the binary nature of the response, one should consider the estimator as the optimizer of a more suited objective function, such as the log-likelihood one, instead of the least squares in (2).

Specifically, given the  $\mathbb{S}^{d-1} \times \{0, 1\}$ -valued random sample  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ , the log-likelihood connected with the binary regression is  $\sum_{i=1}^n \{Y_i \log(\lambda(\mathbf{X}_i)) + (1 - Y_i) \log(1 - \lambda(\mathbf{X}_i))\}$ . The corresponding locally weighted version, at  $\mathbf{x} \in \mathbb{S}^{d-1}$ , can be expressed as

$$\sum_{i=1}^n \left\{ Y_i \log \left( \frac{\lambda(\mathbf{X}_i)}{1 - \lambda(\mathbf{X}_i)} \right) + \log(1 - \lambda(\mathbf{X}_i)) \right\} K_\kappa(\mathbf{x}' \mathbf{X}_i).$$

Setting  $\delta = \log(\lambda/(1 - \lambda))$ , the expression above can be re-written as

$$\sum_{i=1}^n \{Y_i \delta(\mathbf{X}_i) - \log(1 + \exp(\delta(\mathbf{X}_i)))\} K_\kappa(\mathbf{x}' \mathbf{X}_i),$$

and, approximating  $\delta(\mathbf{X}_i)$  around  $\mathbf{x}$  in the local log-likelihood function by using expansion (1), a class of nonparametric estimators for  $\lambda(\mathbf{x})$  can be obtained. Specifically, let  $\beta_0 = \delta(\mathbf{x})$ , and let  $\boldsymbol{\beta}_\ell$  be the matrix of the derivatives of total order  $\ell \in (1, \dots, p)$  of  $\delta$  at  $\mathbf{x}$ . Then, for fixed  $\mathbf{x} \in \mathbb{S}^{d-1}$ , by expressing  $\mathbf{X}_i$  according to the tangent normal decomposition, we define

$$q_p(\mathbf{X}_i; \beta_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) = \beta_0 + \sum_{\ell=1}^p \frac{\theta_i^\ell}{\ell!} \boldsymbol{\xi}_i' \boldsymbol{\beta}_\ell \boldsymbol{\xi}_i^{\otimes(\ell-1)}.$$

Hence, under suitable smoothness assumptions, the  $p$ -degree expansion of the log-likelihood can be expressed as

$$\sum_{i=1}^n \{Y_i q_p(\mathbf{X}_i; \beta_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) - \log(1 + \exp(q_p(\mathbf{X}_i; \beta_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)))\} K_\kappa(\mathbf{x}' \mathbf{X}_i). \quad (4)$$

It is interesting to note that, when  $\kappa$  goes to 0, the kernel  $K_\kappa(\mathbf{x}' \mathbf{X}_i)$  approaches the uniform density and assigns the same weight to each sample point, for any  $\mathbf{x}$ . As a consequence, for  $\kappa$  going to 0, the local log-likelihood optimization reduces to the standard logistic regression problem with spherical predictor.

Now, letting  $\hat{\beta}_0$  be the solution for  $\beta_0$  of the maximization of (4) with respect to  $\{\beta_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p\}$ , a  $p$ -degree local polynomial estimator for  $\lambda(\mathbf{x})$  is

$$\hat{\lambda}_L(\mathbf{x}; p) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)}.$$

When  $p = 0$ , the resulting estimator is the *local constant* one previously discussed, while, when  $p = 1$ , we obtain the spherical version of the *local linear* logistic estimator studied in the Euclidean setting by Fan et al. (1995) and Signorini and Jones (2004). A closed-form expression for  $\hat{\lambda}_L(\mathbf{x}; 1)$  does not exist, but, obviously, distinctly from  $\hat{\lambda}(\mathbf{x}; 1)$ , the estimator always takes value on  $[0, 1]$ . Concerning the asymptotic properties, by following Theorems 3 and 4 of Fan et al. (1995) with  $g$  being the logit link, and by using [Result 1](#), we get the following.

**Result 2.** *Given a  $\mathbb{S}^{d-1} \times \{0, 1\}$ -valued random sample  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ , consider the estimator  $\hat{\lambda}_L(\mathbf{x}; 1)$ ,  $\mathbf{x} \in \mathbb{S}^{d-1}$ . If assumptions (i) and (ii) of [Result 1](#) hold, then we have*

$$\mathbb{E}[\hat{\lambda}_L(\mathbf{x}; 1)] - \lambda(\mathbf{x}) = \frac{b_2(\kappa)}{2(d-1)} \text{Tr} \{ \nabla_{\hat{\delta}}^2(\mathbf{x}) \} \lambda(\mathbf{x}) (1 - \lambda(\mathbf{x})) + o(b_2(\kappa)),$$

$$\text{Var}[\hat{\lambda}_L(\mathbf{x}; 1)] = \frac{v_0(\kappa)}{n} \frac{\lambda(\mathbf{x})(1 - \lambda(\mathbf{x}))}{f(\mathbf{x})} + o\left(\frac{v_0(\kappa)}{n}\right).$$

Notice that  $\hat{\lambda}_L(\mathbf{x}; 1)$  shares both the asymptotic variance and the order of the asymptotic bias of  $\hat{\lambda}(\mathbf{x}; p)$ ,  $p \in (0, 1)$ . Moreover, the asymptotic bias depends only on  $\lambda$  and the derivatives of  $\lambda$  but not on  $f$ , as it happens for  $\hat{\lambda}(\mathbf{x}; 1)$ .

Clearly, by virtue of [Result 2](#), if a von Mises–Fisher kernel is employed as the weight, by recalling the approximations in (3), the estimator attains the convergence rate of order  $n^{-4/(d+3)}$ .

Concerning the selection of  $\kappa$ , a possible way is to start from a least-squares objective function, and choose the value of  $\kappa$  which minimizes

$$\sum_{i=1}^n \left( Y_i - \hat{\lambda}_{-i}(\mathbf{X}_i; \kappa) \right)^2,$$

where  $\hat{\lambda}_{-i}(\mathbf{X}_i; \kappa)$  stands for the estimate of  $\lambda$  at  $\mathbf{X}_i$  with the  $i$ th sample observation removed. A more natural way is to start from the leave-one-out version of the local log-likelihood, i.e. to select the value of  $\kappa$  maximizing

$$\sum_{i=1}^n \left\{ Y_i \log \left( \frac{\hat{\lambda}_{-i}(\mathbf{X}_i; \kappa)}{1 - \hat{\lambda}_{-i}(\mathbf{X}_i; \kappa)} \right) + \log \left( 1 - \hat{\lambda}_{-i}(\mathbf{X}_i; \kappa) \right) \right\}. \quad (5)$$

**Remark 2.** A generalization of the discussed approach arises from considering different weights for *successes* and *failures* in the local log-likelihood expression, i.e.

$$\sum_{i=1}^n Y_i q_p(\mathbf{X}_i; \beta_0, \beta_1, \dots, \beta_p) K_{\kappa_1}(\mathbf{x}' \mathbf{X}_i) - \log(1 + \exp(q_p(\mathbf{X}_i; \beta_0, \beta_1, \dots, \beta_p))) K_{\kappa_2}(\mathbf{x}' \mathbf{X}_i),$$

with  $K_{\kappa_1}$  and  $K_{\kappa_2}$  being spherical kernels giving weight to the observations of the predictor corresponding to  $Y = 1$  and  $Y = 0$ , respectively.

When  $p = 0$ , the solution for  $\beta_0$  of the maximization in the above local log-likelihood function gives the estimator of  $\lambda(\mathbf{x})$  studied by Di Marzio et al. (2019). This latter is defined by using the kernel estimators, say  $\hat{f}_1(\mathbf{x}; \kappa_1)$  and  $\hat{f}_2(\mathbf{x}; \kappa_2)$ , of the spherical densities  $f_1$  and  $f_2$ , respectively, characterizing the distributions of the predictor in the spaces of successes and failures, based on random samples of respective sizes  $n_1$  and  $n_2$ , i.e.

$$\hat{\lambda}(\mathbf{x}; \kappa_1, \kappa_2) = \frac{n_1 \hat{f}_1(\mathbf{x}; \kappa_1)}{n_1 \hat{f}_1(\mathbf{x}; \kappa_1) + n_2 \hat{f}_2(\mathbf{x}; \kappa_2)}. \quad (6)$$

## 5. Experiments

In this section we use simulation experiments to test the performance of the proposed estimator for classification tasks. In particular, we consider the problem of assigning label 0 or 1 to an observation  $\mathbf{x} \in \mathbb{S}^{d-1}$ . To this end, we adopt the rule according to which  $\mathbf{x}$  is assigned to the population with label 1 if the estimate of  $\lambda(\mathbf{x})$  is greater or equal to 0.5.

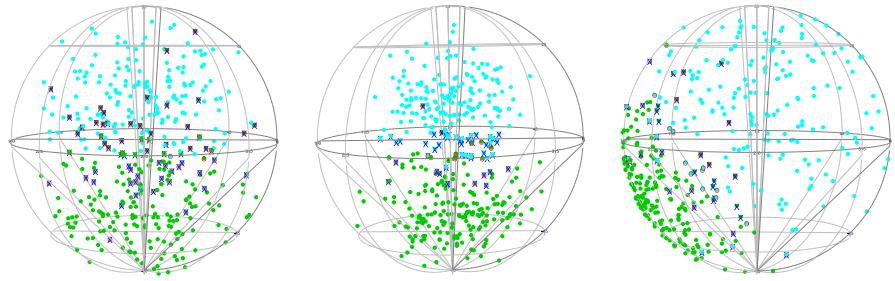
We use  $vMF(\mu, \gamma)$  to denote the von Mises–Fisher distribution on  $\mathbb{S}^2$  with mean direction  $\mu$  (polar co-ordinates expressed in degrees) and concentration parameter  $\gamma$ . We consider different experiments using the following scenarios, where samples of sizes  $n_1 = n_2 = 200$  are, respectively, drawn from  $vMF(\mu_1, \gamma_1)$  and  $vMF(\mu_2, \gamma_2)$ :

- Scenario 1:  $\mu_1 = (270, 20)$ ,  $\mu_2 = (270, -20)$  and  $\gamma_1 = \gamma_2 = 10$ ;
- Scenario 2:  $\mu_1 = (270, 20)$ ,  $\mu_2 = (270, -20)$  and  $\gamma_1 = \gamma_2 = 20$ ;
- Scenario 3:  $\mu_1 = (270, 20)$ ,  $\mu_2 = (220, -20)$ ,  $\gamma_1 = 5$  and  $\gamma_2 = 10$ .

In Scenario 1 the populations, which share the longitude of the mean direction and the value of the concentration parameter, generate rather overlapping groups. Scenario 2 refers to more concentrated populations generating more separated groups. Finally, in Scenario 3 two well-separated groups are generated by populations with different co-ordinates of the mean directions and different concentrations.

As the first experiment, we consider the estimator (6), i.e. the spherical kernel density classifier (SKDC), with  $K_{\kappa_1}$  and  $K_{\kappa_2}$  both being von Mises–Fisher kernels. The smoothing degrees are selected using the von Mises–Fisher reference rule (see Di Marzio et al. 2019 for details). In a second experiment we consider estimator (3). Also in this case we use the von Mises–Fisher kernel as the weight, by selecting the concentration parameter by least squares cross-validation. Fig. 1 illustrates the misclassified observations obtained according to the rule for estimators (3) and (6) by using one dataset for each of the described scenarios. In Table 1, for each experiment, we report as the accuracy measure the average misclassification rate over 200 simulated datasets. The results show that the binary regression estimator slightly outperforms the spherical kernel density classifier, especially when the groups are well-separated. Moreover, the results for  $n_1 = n_2$  show that, in the considered scenarios, estimator (3) performs slightly better than the same estimator using two concentration parameters (which leads to the same classification rule as the kernel density one).

As a real data case study, we apply our methods to the digits dataset used in the StatLog project (Michie et al., 1994). The dataset consists of 18,000 examples of the digits 0 to 9 (i.e.  $q = 10$  classes) extracted from hand-written postcodes in Germany. These numbers were initially digitized onto  $16 \times 16$  images with 256 grey levels. To enable meaningful comparisons with previously obtained results, we have used the same train-test split of the data which has 900 examples of each number (0–9) in the training set and the test set, and an averaging over  $4 \times 4$  pixels resulting in 16 real-valued variables. These data were then transformed to the unit sphere by simply normalizing each observation replacing  $\mathbf{X}_i$  by  $\mathbf{X}_i / \|\mathbf{X}_i\|$ .



**Fig. 1.** From left: Misclassified observations using SKDC (marked by 'X') and using local likelihood with  $p = 0$  (marked by 'o') for one dataset drawn from  $vMF(\mu_1, \gamma_1)$  (cyan points) and  $vMF(\mu_2, \gamma_2)$  (green points) in Scenarios 1, 2 and 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 1**

Estimates of the misclassification rates for SKDC = spherical kernel density classifier and LBR = local binary regression with  $p = 0$ , using 200 samples of sizes  $n_1 = n_2 = 200$ , respectively, drawn from  $f_j = vMF(\mu_j, \gamma_j), j \in (1, 2)$ , given in Scenarios 1–3. For both classification rules we use a von Mises–Fisher kernel: for SKDC,  $\kappa_1$  and  $\kappa_2$  are selected according to the von Mises–Fisher reference rule, and for LBR  $\kappa$  is selected by least squares cross-validation.

Classification rule	Misclassification rate		
	Scenario 1	Scenario 2	Scenario 3
SKDC	0.178	0.090	0.112
LBR	0.147	0.065	0.086

**Table 2**

Misclassification rates for various methods applied to digits dataset (Michie et al., 1994). LBR = local binary regression; MLR = multinomial logistic regression; LDC = linear discriminant classifier; TRC = top rank classifier; SKDC = spherical kernel density classifier; KDC = kernel density classifier.

LBR	MLR	LDC	TRC	SKDC	KDC
0.043	0.086	0.114	0.047	0.039	0.068

**Table 3**

Confusion matrix for local multinomial logistic classifier applied to German handwritten postcode digits. Columns represent true label, and rows the predicted label.

True digit		0	1	2	3	4	5	6	7	8	9
0	853	4	2	0	5	2	2	2	8	2	
1	4	886	3	3	1	1	0	6	0	8	
2	0	0	858	22	1	3	3	1	16	1	
3	0	0	11	839	0	5	0	2	7	14	
4	4	1	2	0	876	1	4	4	0	4	
5	0	0	7	18	1	868	5	0	11	4	
6	8	0	0	0	3	8	880	0	5	0	
7	2	5	0	2	3	0	0	870	1	22	
8	21	2	12	9	1	5	6	4	845	11	
9	8	2	5	7	9	7	0	11	7	834	

Our implementation, which corresponds to a 1-degree local polynomial estimator, used logistic regression with weights obtained from a spherical kernel. The smoothing parameter was selected –for each pair of classes  $(j, k) \in \{0, 1, \dots, 9\} \times \{0, 1, \dots, 9\}$ –using cross-validation (i.e. Eq. (5)), which yielded solutions for the smoothing parameter ranging from 0.9 to 38.8. Then, for each element of the test set, we compute the probability of membership of class  $j$ , given an alternative of class  $k$ , say  $P_{jk}$  with  $P_{jk} = 1 - P_{kj}$  (also setting  $P_{jj} = 1$ ), using the corresponding  $\kappa_{jk} (= \kappa_{kj})$  found by cross-validation. Finally, we allocate this observation to the class  $\text{argmax}_k \min_j P_{jk}$ . The error rate for 9000 observations in the test set is better than those given in Michie et al. (1994, p. 136), i.e. the unweighted multinomial logistic regression, the simple linear discriminant, and the top rank classifier (k-nearest neighbour), see Table 2. The confusion matrix for this classifier is shown in Table 3, in which it can be seen that the most common classification mistakes were to recognize an “8” as a “0”, a “2” as a “3”, and a “7” as a “9”. Although the error rate is very good, we note that this approach was computationally intensive, with the multinomial logistic model entailing the estimation of  $q(q - 1)/2 = 45$  smoothing parameters in the training phase, and a further fitting

of  $nq(q-1)/2 = 405,000$  models in the testing phase. Whilst it would be straightforward to consider  $p = 2$  (including interaction terms, if desired) this would take an excessive amount of time without a common choice of  $\kappa$  across all class pairs.

Using the classification rule based on a kernel density estimator proposed by Di Marzio et al. (2019), a single smoothing parameter (for all classes) was selected by leave-one-out cross-validation on the training data. This value of  $\kappa$  ( $= 140.6$ ) was then used to classify the test data. For this classifier, the error rate was better than the results given in Michie et al. (1994) for data which have not been transformed to the sphere making the transformation onto a non-linear space beneficial, see Table 2.

## Acknowledgements

The authors are grateful to Laura Sangalli, Alessandro Fassò and two anonymous referees for their valuable comments which led to improvements in this article.

## References

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C., 2018. Circular local likelihood. *TEST* 27 (4), 921–945.

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C., 2019. Kernel density classification for spherical data. *Statist. Probab. Lett.* 144, 23–29.

Di Marzio, M., Panzera, A., Taylor, C.C., 2009. Local polynomial regression for circular predictors. *Statist. Probab. Lett.* 79, 2066–2075.

Di Marzio, M., Panzera, A., Taylor, C.C., 2014. Nonparametric regression for spherical data. *J. Amer. Statist. Assoc.* 109, 748–763.

Di Marzio, M., Panzera, A., Taylor, C.C., 2018. Nonparametric rotations for sphere-sphere regression. *J. Amer. Statist. Assoc.* <http://dx.doi.org/10.1080/01621459.2017.1421542>.

Fan, J., Heckman, N., Wand, M.P., 1995. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* 90, 141–150.

Fernandes, K., Cardoso, J.S., 2016. Discriminative directional classifiers. *Neurocomputing* 207, 141–149.

García-Portugués, E., Van Keilegom, I., Crujeiras, R.M., González-Manteiga, W., 2016. Testing parametric models in linear-directional regression. *Scand. J. Statist.* 43, 1178–1191.

Hall, P., Watson, G., Cabrera, J., 1987. Kernel density estimation with spherical data. *Biometrika* 74, 751–762.

Klemela, J., 2000. Estimation of densities and derivatives of densities with directional data. *J. Multivariate Anal.* 73, 18–40.

Ley, C., Verdebout, T., 2017. *Modern Directional Statistics*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Ley, C., Verdebout, T., 2018. *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Mardia, K.V., Jupp, P.E., 2008. *Directional Statistics*. J. Wiley, Chichester.

Michie, D., Spiegelhalter, D.J., Taylor, C.C. (Eds.), 1994. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Chichester.

Signorini, D.F., Jones, M.C., 2004. Kernel estimators for univariate binary regression. *J. Amer. Statist. Assoc.* 99, 119–126.